International Journal of Recent Advances in Multidisciplinary Topics
Volume 2, Issue 6, June 2021
https://www.ijramt.com | ISSN (Online): 2582-7839

Credit Card Fraud Detection Using Isolation Forest

Gaurav Kumar Singh[1], Akhilesh Bhayye[2*], Sanika Dhamnaskar[3], Sandeep Patil[4], S. V. Phulari[5]
[1,2,3,4]*Student, Department of Computer Engineering, PDEA"s College of Engineering, Pune, India*
[5]*Professor, Department of Computer Engineering, PDEA"s College of Engineering, Pune, India*

***Abstract*: Nowadays credit card use has become extremely common. Generally, credit card fraud activity can happen both online and offline. Nowadays most people use online transaction due to which increasing in online transactions by using different payment methods, such as credit/debit card PhonPe, Gpay, Paytm, etc., fraudulent activities have also increased. Credit card fraud stands as a major problem for the world financial institute. According to an RTI report 2480 cases of fraud in 18 public sectors involving Rs. 31, 898, 63. According to RBI in 2017-2018 total 911 credit card fraud amounting to 65.6 crore. The acceptance and rejection of a transaction process happens within a micro or millisecond. Therefore, the detection of a fraud transaction must be extremely quick and effective. There are more than a million transactions which occur daily, and it is difficult to monitor each transaction individually. Thus, an effective fraud detection system is used to differentiate genuine and a fraud transaction. Our project plan to illustrate the design of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes design past credit card transactions with the data of ones that turned out to be fraud. By using this model, we recognize whether a new transaction is fraudulent or not.**

***Keywords*: credit card, credit card fraud detection, machine learning, classification technique, transaction.**

1. Introduction

In today's world as online transaction is increasing. Due to which online fraud has been increasing rapidly all over worldwide. We can avoid fraud by prevention and detection. Prevention avoids any type of attacks from fraudsters. In the case of an online mode of payment, the card may not present physically. In this type of case, the card data is prone to attack by the hacker or cyber-criminal. According to Nilson Report in October 2016, more than $31 trillion were generated worldwide by online payment system in 2015, increasing 7.3% than 2014. The worldwide online fraud detection market is expected to increase at a CAGR of 14.10% during the period of 2019-2025. The purpose of this report is to define, describe, segment, and predict the worldwide online payment fraud detection market on basis of solution, mode, and regions. There are two type of fraud are identified in a set of transactions are Card-not-Present (CNP) fraud and Card-Present (CP) frauds. Card-not-present is type of fraud in which customer does not physically present the card seller during the fraudulent transaction. It is harder to prevent the card-present fraud because the seller cannot personally examine the credit card signs of possible fraud. Our project aims at addressing for fraud natures that belong the CNP fraud category describe. There are many type of frauds are faced during fraud section. The process of whole transaction or rejection are taking place in very small-time interval. Therefor process for detection of fraudulent transaction must be extremely quick and effective. Thus, an efficient Fraud Detection System must be put into work to able to differentiate between genuine and a fraud transaction. The main objective of this paper is to evaluate an imbalance data set with the help of various supervised learning models and find out which one of them is suited for detecting card fraud. We will use 3 supervise learning model for to evaluate a dataset based on various predefined criteria.

2. Machine Learning

Machine learning is a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

Machine Learning can be classified into three types:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

3. Related Work

There are many algorithms that are and have been implemented for this credit card fraud detection. Mostly classification techniques are used such as K-means clustering, KNN, etc. Trees are also used for credit card fraud detection i.e., Decision Tree and Random Forest. But many a times Decision Trees led to overfitting and high computational cost, so to overcome this anomaly detection techniques are used like Isolation Forest.

4. Dataset

Dataset contains data that is related to each other i.e., related data. For this paper we have made use of a publicly available dataset. This dataset contains the record of transactions made

*Corresponding author: ab.bhayye@gmail.com

by European cardholders. It has the records of 284,807 transactions made over a span of two days, out of which 492 were found out to be fraud. There is total 31 variables in the dataset including the dependent one. This is an imbalance dataset. Imbalance dataset means there is an unequal distribution of classes in the dependent variable.

5. Anomaly Detection

One of the use case of anomaly detection is outlier's detection. Predictive maintenance usually takes anomaly detection because it is much more difficult to learn what defects is and what is default or normal. Also used in health care and financial frauds and network security. Mapping the features space into a one-dimensional anomaly square, the next step is then to have a very easy model that turns this anomaly square into classifier. Where we can say whether it is an anomaly or not. In this case there are bunch of training samples and prediction that is done on the same data but sometimes you can have anomalies. The training set can also have some anomalies sometimes. It is very similar to classification.

Anomaly=Outlier=Deviant or Unusual Data Point.

When data generating process behaves unusually it results in outliers. Often, the real challenge in anomaly detection is to construct the right data model to separate outliers from noise and normal data.

6. Isolation Forest

Isolation forest is an ensemble regressor and it uses the concept of isolation to separate or classify anomalies. No profiling of normal instances and no point-based distance calculation. If builds an ensemble of random trees for a given data set, and anomalies are points with the shortest average path length. Isolation forest was introduced in 2008 and became available in sck it-learn in 2016. Isolation Forest extends Base Bagging (Bootstrap Aggregated Regressor), and it is possible to control bootstrap parameter (True = will replacement, False = without). Its base estimator is Extra Tree Regressor – an extremely randomized tree regressor. It splits on the best split among randomly chosen attributes with randomly chosen split points. This helps to overcome overfitting and locally greedy trees. Isolation Forest can work as supervised and unsupervised classifier. Isolation forest calculates an anomaly score,

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Where, h(x) is the number of edges in a tree for point x, and c (n) is normalization constant for a data set of size n. From each observed anomaly score following decisions can be taken: If the score is close to one indicates anomaly. If the score is much smaller than 0.5 then it is a normal observation. If the score is close to 0.5 then the sample does not seem to have clearly distinct anomalies.

7. Conclusion

1. Machine Learning can efficiently support fraud detection. It allows to automatize detection and reaction to frauds.
2. Efficient credit card fraud detection system is an utmost requirement for any card issuing bank.
3. In this project we used an imbalanced dataset to check the suitability of different supervised machine learning models to predict the occurrence of a fraudulent transaction.
4. Decision parameters such as sensitivity, time and precision are used.
5. As we are working on imbalanced dataset accuracy parameter is not used because it is not sensitive to imbalanced data.
6. KNN, Decision Tree, Logistic Regression and Random Forest models are used in this study.
7. The sensitivity of the KNN model is greater than that of Decision tree, but as time taken by KNN for testing the data is very large, so we choose Decision Tree over KNN.
8. Decision Tree take minimum time for fraud detection prediction, therefore, Decision Tree is the preferred model.
9. Feasibility of credit card fraud detection based on outlier mining.
10. Outlier detection mining based on distance sum into credit card fraud detection.

References

[1] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi and G. Bontempi," Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, Aug. 2018.

[2] S. Dhankhad, E. Mohammed and B. Far, ' Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study," 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 122-125, 2018.

[3] IOSR Journal of Computer Engineering (IOSR-JCE) vol. 21, no. 3, pp. 45-52, 2019.